Cuestionario de teoría 2

Antonio Álvarez Caballero
analca3@correo.ugr.es

14 de mayo de 2016

1. Cuestiones

Cuestión 1. Sean x e y dos vectores de observaciones de tamaño N. Sea

$$\text{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

la covarianza de dichos vectores, donde $\bar{z}$ representa el valor medio de los elementos de z. Considere ahora una matriz X cuyas columnas representan vectores de observaciones. La matriz de covarianzas asociada a la matriz X es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Defina la expresión matricial que expresa la matriz cov(X) en función de la matriz X.

Solución. ss

Cuestión 2. Considerar la matriz hat definida en regresión, $\mathrm{H} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, donde X es una matriz $N \times (d+1)$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es invertible.

1. Mostrar que H es simétrica
2. Mostrar que $\mathrm{H}^{\mathrm{K}} = \mathrm{H}$ para cualquier entero K

Solución. Resolvamos ambas partes:

1. Para ver que H es simétrica, no hay más que ver que es igual a su traspuesta. Hay que tener en cuenta que la traspuesta cambia el orden del producto matricial, además de que sabemos que $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es simétrica.

$$H^T = \left(\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\right)^T = \mathrm{X}^{\mathrm{T}^{\mathrm{T}}}\left(\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\right)^T = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-\mathrm{T}}\mathrm{X}^{\mathrm{T}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} = H$$

2. Para ver que $H^K = K\ \forall K \in \mathbb{N}$, lo haremos por inducción.

   $\underline{K = 2}$: Es claro que $H^2 = H$:

$$H^2 = X\underbrace{(X^TX)^{-1}X^TX}_{I}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

   $\underline{K \overset{?}{\Rightarrow} K+1}$: Suponiendo que es cierto para K:

$$H^{K+1} = H \Leftrightarrow \underbrace{H^K}_{Inducción} H = H \Leftrightarrow HH = H \Leftrightarrow \underbrace{H^2}_{K=2} = H$$

   Lo cual ya hemos demostrado. Así ya lo hemos probado $\forall K \in \mathbb{N}$

Cuestión 3. Resolver el siguiente problema: Encontrar el punto (x_0, y_0) sobre la línea $ax + by + d = 0$ que este más cerca del punto (x_1, y_1).

Solución. La función a minimizar es la función distancia, que en el plano $\mathbb{R}^2$ está definida por:

$$d(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Como la raíz cuadrada es una función creciente, podemos minimizar lo que queda dentro de la raíz, ya que alcanzan el mínimo en el mismo punto. Por tanto, nuestra función f a minimizar es:

$$f(x, y) = (x_1 - x_2)^2 + (y_1 - y_2)^2$$

Veamos ahora la restricción. El punto (x_0, y_0) que buscamos debe pertenecer a la recta dada luego la restricción g viene dada por:

$$g(x_0, y_0) = ax_0 + by_0 + d$$

Uniendo ambas expresiones, utilizando el método de los multiplicadores de *Lagrange*, nos queda:

$$\mathcal{L}(x_0, y_0, \lambda) = (x_1 - x_0)^2 + (y_1 - y_0)^2 - \lambda(ax_0 + by_0 + d)$$

Ahora debemos calcular el gradiente de este campo escalar, igualar al vector $\vec{0}$ y resolver el sistema de ecuaciones resultante.

$$\frac{\partial \mathcal{L}}{\partial x_0} = 0 \Leftrightarrow -2(x_1 - x_0) + \lambda a = 0 \Leftrightarrow x_0 = x_1 - a\,\frac{\lambda}{2}$$

$$\frac{\partial \mathcal{L}}{\partial y_0} = 0 \Leftrightarrow -2(y_1 - y_0) + \lambda a = 0 \Leftrightarrow y_0 = y_1 - b\,\frac{\lambda}{2}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \Leftrightarrow ax_0 + by_0 + d = 0 \Leftrightarrow a\left(x_1 - \frac{\lambda a}{2}\right) + b\left(y_1 - \frac{\lambda b}{2}\right) + d = 0 \Leftrightarrow \lambda = \frac{2ax_1 + 2by_1 + 2d}{a^2 + b^2}$$

Ahora sustituimos λ en las equivalencias que hemos obtenido para x_0 e y_0:

$$x_0 = x_1 - a\,\frac{ax_1 + by_1 + d}{a^2 + b^2}$$

$$y_0 = y_1 - b\,\frac{ax_1 + by_1 + d}{a^2 + b^2}$$

Por tanto, el punto (x_0, y_0) de la recta $ax + by + d = 0$ más cercano a un punto del plano (x_1, y_1) es

$$(x_0, y_0) = (x_1 - a\,\frac{ax_1 + by_1 + d}{a^2 + b^2}, y_1 - b\,\frac{ax_1 + by_1 + d}{a^2 + b^2})$$

Cuestión 4. Consideremos el problema de optimización lineal con restricciones definido por

$$\mathrm{Min}_z \mathrm{c}^{\mathrm{T}}\mathrm{z}$$
$$\text{Sujeto a } \mathrm{Az} \leq \mathrm{b}$$

donde c y b son vectores y A es una matriz.

1. Para un conjunto de datos linealmente separable mostrar que para algún w se debe de verificar la condición $\mathrm{y_n w^T x_n} > 0$ para todo $(\mathrm{x_n}, \mathrm{y_n})$ del conjunto.

2. Formular un problema de programación lineal que resuelva el problema de la búsqueda del hiperplano separador. Es decir, identifique quienes son A, **z**, **b** y **c** para este caso.

Solución. d

Cuestión 5. Probar que en el caso general de funciones con ruido se verifica que $\mathbb{E}_{\mathcal{D}}[E_{out}] = \sigma^2 + \texttt{bias} + \texttt{var}$ (ver transparencias de clase).

Solución. Nuestro problema se reduce a calcular, tal como se hizo en clase, la expresión $\mathbb{E}_{\mathcal{D}}[E_{out}(g^D(x))]$, donde la función objetivo f tiene ruido. Suponemos, como siempre, que el ruido ϵ tiene media 0 y varianza σ^2.

$$\mathbb{E}_{\mathcal{D}}\left[E_{out}\left(g^D(x)\right)\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_x\left[\left(g^D(x) - y(x)\right)^2\right]\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_x\left[\left(g^D(x) - f(x) - \epsilon(x)\right)^2\right]\right]$$

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_x\left[\left(g^D(x) - f(x) - \epsilon(x)\right)^2\right]\right] = \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^D(x) - f(x) - \epsilon(x)\right)^2\right]\right]$$

Desarrollando el trinomio queda:

$$\mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[g^D(x)^2 + f(x)^2 + \epsilon(x)^2 - 2f(x)g^D(x) - 2g^D(x)\epsilon(x) - 2f(x)\epsilon(x)\right]\right]$$

Ahora sumamos y restamos $\bar{g}(x)^2$ para obtener las expresiones para $bias(x)$ y $var(x)$.

$$\mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[g^D(x)^2 - \bar{g}(x)^2 + \bar{g}(x)^2 + f(x)^2 + \epsilon(x)^2 - 2f(x)g^D(x) - 2g^D(x)\epsilon(x) - 2f(x)\epsilon(x)\right]\right]$$

Usamos la linealidad de la esperanza y reordenamos para obtener dichas expresiones. La varianza es la indicada porque el doble producto de que debe salir del binomio al cuadrado es 0, ya que $\bar{g}(x)$ no depende de $\mathcal{D}$.

$$\mathbb{E}_x\left[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^D(x)^2\right] - \bar{g}(x)^2}_{var(x)} \underbrace{+\bar{g}(x)^2 - 2f(x)\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^D(x)\right]}_{\bar{g}(x)} + f(x)^2 - 2\epsilon(x)\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^D(x)\right]}_{\bar{g}(x)} - 2f(x)\epsilon(x) + \epsilon(x)^2}_{bias(x)}\right]$$

Entonces nos queda

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^D(x))] = \mathbb{E}_x\left[var(x) + bias(x) - 2\epsilon(x)\bar{g}(x) - 2f(x)\epsilon(x) + \epsilon(x)^2\right]$$

Ahora usamos de nuevo la linealidad de la esperanza. Además, destacar que $f(x)$ y $\epsilon(x)$ son independientes, luego la esperanza del producto es el producto de las esperanzas.

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^D(x))] = var + bias - 2\underbrace{\mathbb{E}_x\left[\epsilon(x)\right]}_{0}\mathbb{E}_x\left[\bar{g}(x)\right] - 2\mathbb{E}_x\left[f(x)\right]\underbrace{\mathbb{E}_x\left[\epsilon(x)\right]}_{0} + \underbrace{\mathbb{E}_x\left[\epsilon(x)^2\right]}_{\sigma^2}$$

Con esto, ya lo tenemos: por hipótesis, sabemos que la media del ruido es 0 y la varianza σ^2, luego nos queda la expresión que buscábamos:

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^D(x))] = var + bias + \sigma^2$$

Cuestión 6. Consideremos las mismas condiciones generales del enunciado del Ejercicio.2 del apartado de Regresión de la relación de ejercicios.2. Considerar ahora $\sigma = 0{,}1$ y $d = 8$, ¿cual es el más pequeño tamaño muestral que resultará en un valor esperado de E_{in} mayor de 0,008?.

Solución. Conociendo la expresión del valor esperado del E_{in} y acotándolo inferiormente por 0,008:

$$\mathbb{E}_{\mathcal{D}}\left[E_{in}(w_{lim})\right] = \sigma^2\left(1 - \frac{d+1}{N}\right) \geq 0{,}008$$

Ahora sólo sustituimos los valores que nos han dado y despejamos el tamaño de la muestra N.

$$0{,}1^2\left(1 - \frac{9}{N}\right) \geq 0{,}008 \Rightarrow \left(1 - \frac{9}{N}\right) \geq 0{,}8 \Rightarrow -\frac{9}{N} \geq -0{,}2 \Rightarrow \frac{9}{0{,}2} \leq N \Rightarrow 45 \leq N$$

Así, el mínimo N para que el valor esperado de E_{in} sea 0,008 es 45.

Cuestión 7. En regresión logística mostrar que

$$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1 + e^{y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n)$$

Argumentar que un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado.

Solución. Partimos de la expresión de E_{in} conocida:

$$E_{in}(w) = \frac{1}{N}\sum_{n=0}^{N}\ln\left(1 + e^{-y_i w^T x_i}\right)$$

Ahora sólo calculamos su gradiente con respecto a w, por lo que coincide con su parcial con w.

$$\nabla E_{in}(w) = \frac{\partial}{\partial x}\left(\frac{1}{N}\sum_{n=0}^{N}\ln\left(1 + e^{-y_i w^T x_i}\right)\right) = \frac{1}{N}\sum_{n=1}^{N}\frac{-y_n x_n e^{-y_n w^T x_n}}{1 + e^{-y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N}\frac{-y_n x_n}{1 + e^{y_n w^T x_n}}$$

$$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1 + e^{y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n)$$

Es claro que una solución mal clasificada influye más que una buena. Una solución bien clasificada equivale a $y_n w^T x_n > 0$, luego la exponencial del denominador será mayor que 1. Al dividir por algo más grande, el cociente se hace más pequeño. En el caso de $y_n w^T x_n < 0$, la exponencial será una exponencial negativa, con valores menores a 1. Luego el denominador será más pequeño, y el cociente más grande. Luego la sumatoria será más grande.

Cuestión 8. Definamos el error en un punto (x_n, y_n) por

$$e_n(w) = \max(0, -y_n w^T x_n)$$

Argumentar que el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $\eta = 1$.

Solución. Recordamos que la función de actualización de los pesos para el SGD es

$$w = w - \eta \cdot \nabla e_n(w)$$

Además, la función de error está bien definida. Si un dato está mal clasificado, $y_i w^T x_i > 0$, por lo cual el máximo de 0 y $-y_i w^T x_i$ es 0. En caso de estar mal clasificado, el máximo es dicho valor.

Vamos a calcular el gradiente del error y veamos qué sale.

$$\nabla e_n(w) = \frac{\partial}{\partial w}\Big(max(0, -y_n w^T x_n)\Big) = max(0, -y_n x_n)$$

Entonces, para datos mal clasificados, la función de actualización de los pesos es:

$$w = w - \eta \cdot (-y_n x_n) \overset{\eta=1}{\Rightarrow} w = w + 1 \cdot y_n x_n$$

Que es justo la fórmula para actualizar pesos en el PLA.

Cuestión 9. El ruido determinista depende de $\mathcal{H}$, ya que algunos modelos aproximan mejor f que otros.

1. Suponer que $\mathcal{H}$ es fija y que incrementamos la complejidad de f.
2. Suponer que f es fija y decrementamos la complejidad de $\mathcal{H}$

Contestar para ambos escenarios: ¿En general subirá o bajará el ruido determinista? ¿La tendencia a sobrejaustar será mayor o menor? (Ayuda: analizar los detalles que influencian el sobreajuste)

Solución. dfasdf

Cuestión 10. La técnica de regularización de Tikhonov es bastante general al usar la condición

$$w^T \Gamma^{\mathrm{T}} \Gamma w \leq C$$

que define relaciones entre las w_i (La matriz Γ_i se denomina regularizador de Tikhonov)

1. Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$
2. Calcular Γ cuando $(\sum_{q=0}^{Q} w_q)^2 \leq C$

Argumentar si el estudio de los regularizadores de Tikhonov puede hacerse a través de las propiedades algebraicas de las matrices Γ.

Solución. asdf

BONUS. Considerar la matriz $\hat{\mathrm{H}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$. Sea X una matriz $N \times (d+1)$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ invertible. Mostrar que traza$(\hat{\mathrm{H}}) = d + 1$, donde traza significa la suma de los elementos de la diagonal principal. (+1 punto)

Solución. Para resolver esto sólo necesitamos utilizar una propiedad básica de álgebra lineal:

$$\mathrm{Traza(AB)} = \mathrm{Traza(BA)}$$

Utilizando esto, dividimos $\hat{H}$ en dos matrices A y B.

$$A = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$$

$$B = \mathrm{X}^{\mathrm{T}}$$

Es claro que $AB = \hat{H}$. Es claro que BA es la matriz identidad. Lo único que hay que tener cuidado es con las dimensiones de esta matriz.

$$\mathrm{X} \in \mathcal{M}_{N \times d+1} \Rightarrow \mathrm{X^T X} \in \mathcal{M}_{d+1 \times d+1}$$

Como una matriz y su inversa tienen la misma dimensión, entonces:

$$BA = \underbrace{X^T X}_{\mathrm{M}} \underbrace{(X^T X)^{-1}}_{\mathrm{M^{-1}}} = I_{d+1}$$

Y la traza de esta matriz es $d + 1$. Como coincide con la traza de $\hat{H}$, podemos concluir que

$$\mathrm{traza(\hat{H})} = d + 1$$